European Journal of Human Genetics (2015) 23, 22–28

www.nature.com/ejhg

ARTICLE

Toward a common language for biobanking

Martin N Fransson*[,1,2], Emmanuelle Rial-Sebbag[3], Mathias Brochhausen[4] and Jan-Eric Litton[1]

To encourage the process of harmonization, the biobank community should support and use a common terminology. Relevant terms may be found in general thesauri for medicine, legal instruments or specific glossaries for biobanking. A comparison of the use of these sources has so far not been conducted and would be a useful instrument to further promote harmonization and data sharing. Thus, the purpose of the present study was to investigate the preference of definitions important for sharing biological samples and data. Definitions for 10 terms –*[human] biobank*, *sample*/*specimen*, *sample collection*, *study*, *aliquot*, *coded*, *identifying information*, *anonymised*, *personal data* and *informed consent*–were collected from several sources. A web-based questionnaire was sent to 560 European individuals working with biobanks asking to select their preferred definition for the terms. A total of 123 people participated in the survey, giving a response rate of 23%. The result was evaluated from four aspects: scope of definitions, potential regional differences, differences in semantics and definitions in the context of ontologies, guided by comments from responders. Indicative from the survey is the risk of focusing only on the research aspect of biobanking in definitions. Hence, it is recommended that important terms should be formulated in such a way that all areas of biobanking are covered to improve the bridges between research and clinical application. Since several of the terms investigated here within can also be found in a legal context, which may differ between countries, establishing what is a proper definition on how it adheres to law is also crucial.
European Journal of Human Genetics (2015) **23,** 22–28; doi:10.1038/ejhg.2014.45; published online 9 April 2014

INTRODUCTION

Human biological samples are widely used in clinical trials, observational studies and personalized medicine. Their value can be expected to increase even more in the coming years, as next-generation sequencing technologies will bring forth omics data from sample derivatives at a faster pace.[1,2] Consequently, the importance of well-organized and well-maintained storage facilities for the biological samples with the possibility to compare specimens across different storage facilities, or *biobanks*, should be given a high priority. The European and global biobank community is currently in the process of establishing common infrastructures to promote harmonization[3,4] to make visible both samples and data and provide a standardized way for sharing these resources. Already, sharing high-level information about the organizational structure of biobanks and non-sensitive data about the stored samples is the focus of several national and international initiatives.[5,6]

To struggle against barriers to data sharing and to encourage the process of harmonization,[7] the biobank community should support and use a common terminology. Owing to its nature of dealing with both biological samples and potentially sensitive data, the field of biobanking relates to several knowledge domains; biology, to describe the properties of a sample; medicine to annotate associated clinical information; computer science for management of sample data; and law to provide the framework for donor-informed consent and control of personal data. Hence, terms relevant for the biobank community may be found both in general thesauri for medicine and biology[8,9] or legal instruments.[10] In addition, specific glossaries for biobanking have also been developed by regional and international organizations, already recognized for supporting harmonization efforts in the biobank community.[4,11–13] To our knowledge, a survey comparing the use of these sources has so far not been conducted, and would be a useful instrument to further promote harmonization and data sharing. Thus, the purpose of the present study was to investigate the preference of definitions for 10 terms often used in biobanking. We have used questionnaires aiming to answer which definition is the preferred one for each of the terms, with a succeeding discussion guided by the quantitative result and comments from responders.

MATERIALS AND METHODS

Ten terms were selected on the basis of being important for information sharing about biological samples – for instance, in the implementation of a query system. For such systems, the first five terms, described in Table 1, are often used in an explicit way as data variables or attributes, describing *what information and samples are being shared*. On the other hand, the last five terms, described in Table 2, are highly relevant for the process of sharing, describing *the conditions for sharing information about samples*. The selected terms were [HUMAN] BIOBANK, SAMPLE and/or SPECIMEN, SAMPLE COLLECTION, STUDY, ALIQUOT, CODED/CODING, IDENTIFYING INFORMATION/IDENTIFIABILITY, ANONYMISED/ANONYMISATION, PERSONAL DATA and INFORMED CONSENT. Definitions were collected from P[3]G,[11] ISBER,[12] OECD,[13,14] Medical Subject Headings (MeSH),[8] Statutes for the Biobanking and Biomolecular Resources Research Infrastructure — European Research Infrastructure Consortium (BBMRI-ERIC),[15] The National Cancer Institute Thesaurus (NCI),[9] the German Ethics Council,[16] the Swedish Association of Local Authorities and Regions,[17] the Oxford English Dictionary[18] and the Directive 95/46/EC of the European Parliament and of the Council.[10] The questionnaire emphasized that all terms should be considered in the context of biobanks.

[1]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [2]Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden; [3]INSERM/Université de Toulouse—Université Paul Sabatier, Toulouse III, UMR 1027, Toulouse, France; [4]Division of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA
*Correspondence: Dr MN Fransson, Institute of Environmental Medicine, Karolinska Institutet, PO Box 210, Stockholm SE-171 77, Sweden. Tel: +46 8 524 822 30; Fax: +46 8 33 69 81; E-mail: martin.fransson@ki.se
Received 22 August 2013; revised 29 January 2014; accepted 19 February 2014; published online 9 April 2014

Table 1 Definitions for [HUMAN] BIOBANK, SAMPLE and/or SPECIMEN, SAMPLE COLLECTION, STUDY and ALIQUOT

Term and definition	*Source*
[HUMAN] BIOBANK	
Collections, repositories and distribution centres of all types of human biological samples, such as blood, tissues, cells or DNA and/or related data such as associated clinical and research data, as well as biomolecular resources, including model- and microorganisms that might contribute to the understanding of the physiology and diseases of humans	BBMRI-ERIC
An organized collection of human biological material and associated information stored for one or more research purposes	P^3G
Facilities that collect, store and distribute tissues – e.g., cell lines, microorganisms, blood, sperm, milk, breast tissue, for use by others. Other uses may include transplantation and comparison of diseased tissues in the identification of cancer	MeSH[a]
A material entity consisting of storage facilities for specimens (DNA, blood, tissue) derived from humans and information related to these specimens	German Ethics Council[b]
An entity that receives, stores, processes and/or disseminates specimens, as needed. It encompasses the physical location as well as the full range of activities associated with its operation	ISBER
SAMPLE and/or SPECIMEN	
[Sample] A single unit containing material derived from one specimen AND [Specimen] A specific tissue, blood sample, urine sample and so on, obtained from a single participant at a specific time	OECD (2009)
[Sample] A biological specimen from the human body including – e.g., tissue, blood, blood components, cell lines and biopsies	P^3G
[Biospecimen] Any material sample taken from a biological entity for testing, diagnostic, propagation, treatment or research purposes, including a sample obtained from a living organism or taken from the biological object after halting of all its life functions. Biospecimen can contain one or more components including but not limited to cellular molecules, cells, tissues, organs, body fluids, embryos and body excretory products	NCI
[Biological sample] A biological specimen including – for example, blood, tissue, urine, and so on taken from a participant	OECD (2006)
[Sample] A single unit containing material derived from one specimen AND [Specimen] A specific tissue, blood sample and so on taken from a single subject or donor at a specific time	ISBER
SAMPLE COLLECTION	
A collection of samples with at least one common characteristic	Swedish Association of Local Authorities and Regions
A number of samples collected or gathered together, viewed as a whole	Oxford English Dictionary[c]
A group of samples that has been isolated for future research purposes	ISBER
STUDY	
Studies designed to examine associations, commonly, hypothesized causal relations. They are usually concerned with identifying or measuring the effects of risk factors or exposures	MeSH[d]
A detailed examination, analysis or critical inspection of a subject designed to discover facts about it	NCI
ALIQUOT	
A portion of a sample of biological material that has been divided into separated parts	P^3G
Pertaining to a portion of the whole; any one of two or more samples of something, of the same volume or weight	NCI
A process wherein a specimen is divided into separate parts that are typically stored in separate containers as individual samples. The term aliquot may also be used as a noun to denote a single sample	ISBER

Abbreviations: BBMRI-ERIC, Biobanking and Biomolecular Resources Research Infrastructure—European Research Infrastructure Consortium; MeSH, Medical Subject Headings; NCI, The National Cancer Institute Thesaurus.
[a]Original entry is [Biological Specimen Banks].
[b]Adapted from: 'Human biobanks usually refer to collections of samples of human body substances (eg, tissue, blood, DNA), which are linked to personal data and sociodemographic information about the donors of the material.'
[c]Adapted from [Collection].
[d]Original entry is [Epidemiologic Studies]. Original definition also included: 'The common types of analytic study are CASE–CONTROL STUDIES; COHORT STUDIES; and CROSS-SECTIONAL STUDIES.'

The questionnaire was designed using Websurvey (Textalk, Mölndal, Sweden) with a predefined set of two to five definitions for each term, depending on the number of relevant definitions that could be found in literature. In addition, as an alternative, the respondent could choose to enter a comment.

For SAMPLE COLLECTION, only two definitions could be found in literature from sources relating to biobanks; one from the Swedish Association of Local Authorities and Regions and one from ISBER. To create more alternatives, the definition of COLLECTION from the Oxford English Dictionary was adapted and used in this context. Of the three definitions only the one from the Swedish Association of Local Authorities and Regions explicitly defines how the samples in a collection are related, with at least one common characteristic. Proper definitions for STUDY in the context of biobanks were, similar to SAMPLE COLLECTION, difficult to find in literature. Of the two given definitions, the one by MeSH defines STUDY in the context of epidemiology, whereas the definition given by NCI is generic. The selection thus offered two contrasting definitions, where the use of the first one can be motivated by the fact that epidemiology is a research field highly linked with biobanking. In some cases, sources did not use semantically or syntactically identical terms; for example, CODED *vs* CODING, but as the definitions were not strict in the same sense these terms were lumped for comparison.

To avoid, as far as possible, bias caused by the responders being more familiar with a particular organization, no sources were included in the questionnaire, and the definitions were also not put in a particular order.

The questionnaire was sent by e-mail to an European group ($N = 438$), comprising one or two contact persons per biobank or biobank network in the Catalog of European Biobanks,[5] and to a Swedish group ($N = 122$), according

Table 2 Definitions for CODED/CODING, IDENTIFYING INFORMATION/IDENTIFIABILITY, ANONYMISED/ANONYMISATION, PERSONAL DATA and INFORMED CONSENT

Term and definition	*Source*
CODED/CODING	
Where data and samples are labelled with at least one specific code and do not carry any personal identifiers	OECD (2009)[a]
Substituting a code for personally identifying information in such a way that linkage is only possible through a key	P^3G[b]
IDENTIFYING INFORMATION/IDENTIFIABILITY	
Information that may lead to the identification of the participant from whom the human biological material, data and associated information are obtained	OECD (2009)[c]
Any combination of data that allows a specific person to be identified. There are various terms used to describe this (eg, coding [single or double], linkage, traceability, pseudonyminization and so on).	P^3G[d]
Information (Eg, name, social security number, medical record or pathology accession number and so on.) that would enable the identification of the subject. For some specimens this information might include the taxon name and collection number	ISBER[e]
ANONYMISED/ANONYMISATION	
Anonymised data and samples are initially single or double coded but where the link between the subjects' identifiers and the unique code(s) is subsequently deleted. Once the link has been deleted, it is no longer possible to trace the data and samples back to individual subjects through the coding key(s)	OECD (2009)[f]
The irreversible removal of personal identifiers from data or samples, such that no specific individual can be identified	P^3G[g]
PERSONAL DATA	
Any information relating to an identified or identifiable natural person ('data subject')	Directive 95/46/EC[h]
Any information that directly or indirectly identifies a specific individual	P^3G
INFORMED CONSENT	
A process by which information concerning the intended research is provided to the participant or participant's substitute decisionmaker with an opportunity for them to ask questions, after which specific approval is documented	OECD (2009)
Voluntary authorization, by a patient or research subject, with full comprehension of the risks involved, for diagnostic or investigative procedures and for medical and surgical treatment	MeSH
Voluntary and informed expression of the will of a person, or his/her legal representative, concerning the use(s) of their samples and data. Depending on the nature of the biobank, such consent can take various forms (eg, broad, specific, implicit, proxy, reconsent and so on)	P^3G
A decision to participate in research, taken by a competent individual who has received the necessary information; who has adequately understood the information; and who, after considering the information, has arrived at a decision without having been subjected to coercion, undue influence, inducement or intimidation	ISBER

Abbreviation: MeSH, Medical Subject Headings.
[a]Original entry is [Coded].
[b]Original entry is [Coding].
[c]Original entry is [Identifying information].
[d]Original entry is [Identifiability].
[e]Original entry is [Identifier/Identifying information].
[f]Original entry is [Anonymised/Anonymisation]. Original definition also included: 'Anonymisation is intended to prevent subject re-identification. As anonymised samples and associated data are not traceable back to the subject, it is not possible to undertake actions such as sample withdrawal, or the return of individual results, even at the subject's request. The use of anonymised data and samples does not allow for clinical monitoring, subject follow-up or the addition of new data from the subject. The deletion of the coding key(s) linking the data and samples to a given subject's identifiers provides additional confidentiality and privacy protection over coded data and samples, as it prevents subject re-identification through the use of the coding key(s).'
[g]Original entry is [Anonymization].
[h]Original definition also included: 'An identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity.'

to a Swedish e-mailing list for biobanks and registries. An English, respective Swedish, cover letter and header were sent with the questionnaires, which were otherwise identical. The survey period lasted from 28 June to 7 September 2012, with two reminders after 1 and 2 months.

RESULTS

Of the 438 European biobank contacts, 92 responded, giving a response rate of 22%, if also considering that the e-mail was permanently undeliverable to 21 addresses. The 'type of biobank', as classified in the Catalog of European Biobanks (1) 'Clinical biobank/study', (2) 'Population-based biobank/study' or (3) 'Non-human biobank/study' was retrieved for the 92 responders. Fourteen responders were pairwise affiliated to the same biobank organization. Eight responders were affiliated to a network of biobanks rather than a specific organization, and one responder could not be traced back to a particular biobank organization or network. The distribution of European responders among clinical, population-based and biobank networks are presented in Figure 1. In a similar manner, the 'country of biobank' for each responder was retrieved from the Catalog of European Biobanks. The country of affiliation for the 346 non-responders was determined using the country domain of their respective e-mail address, or retrieved from the Catalog of European Biobanks for biobanks and networks categorized as EU or when a country domain was not part of the e-mail address. The number of responders versus invited participants for each country is presented in Figure 2. Invited participants with permanently undeliverable e-mail addresses have been excluded.

In the Swedish group, 31 out of 122 responded, giving a response rate of 25%. A retrospective categorization by affiliation of type of biobank was not possible for the Swedish respondents. Taken together (All), 123 people participated in the survey, giving a total response rate of 23%. The results for each term are presented in Tables 3 and 4.

[HUMAN] BIOBANK

Of the five definitions for [HUMAN] BIOBANK the one by P^{3}G got the highest rating, although closely followed by the definition used in the BBMRI-ERIC statutes. Four of the European respondents chose to enter a comment instead of selecting one of the specified definitions. Three respondents made a reference to the definitions of EuroBioBank,[19] the Marble Arch International Working Group on Biobanking for Biomedical Research[20] and the Norwegian body of law.[21] One respondent emphasized that the clinical use of biobanks should also be part of a definition.

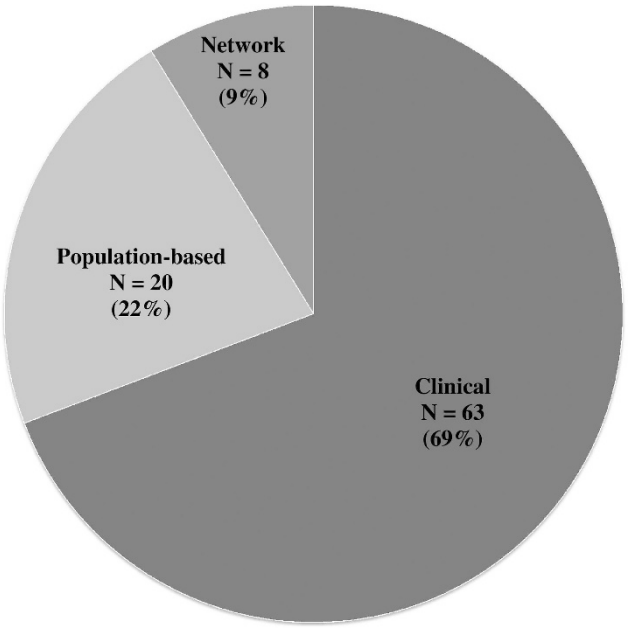

Figure 1 Type of biobank affiliated with responders from the European group using the classification of clinical and population-based biobanks according to the Catalog of European Biobanks, with the addition of responders who are affiliated to a biobank network instead of a specific biobank.

SAMPLE and/or SPECIMEN

The most popular definition for SAMPLE and/or SPECIMEN was the one issued by the P^{3}G consortium.[11] One of the Swedish respondents chose to enter a comment, suggesting that SAMPLE and SPECIMEN are two different concepts, and that SPECIMEN seems to imply a sample from a sample.

SAMPLE COLLECTION

For the term SAMPLE COLLECTION, the definition by the Swedish Association of Local Authorities and Regions[17] received the highest score. One European respondent chose to enter a comment regarding the definition by ISBER[12] and the interpretation of the term 'isolated' in this definition.

STUDY

Of the two definitions for STUDY, the more general definition by NCI[9] was favored over the definition of STUDY in the epidemiological context provided by MeSH.[8] Two European respondents and one Swedish commented that neither of the definitions are correct, or that they are too narrow, or that biobanks should be regarded as a service for studies and that the concept of STUDY should not be related *per se*. One European respondent commented that biobanks can serve various types of research but can also be used for diagnosis.

ALIQUOT

Of all terms, ALIQUOT received the best consensus among respondents, where the P^{3}G definition[11] was favored in all groups. One of the Swedish respondents made a comment that ALIQUOT corresponds to a sample from a sample according to the Glossary by the Swedish Association of Local Authorities and Regions.[17]

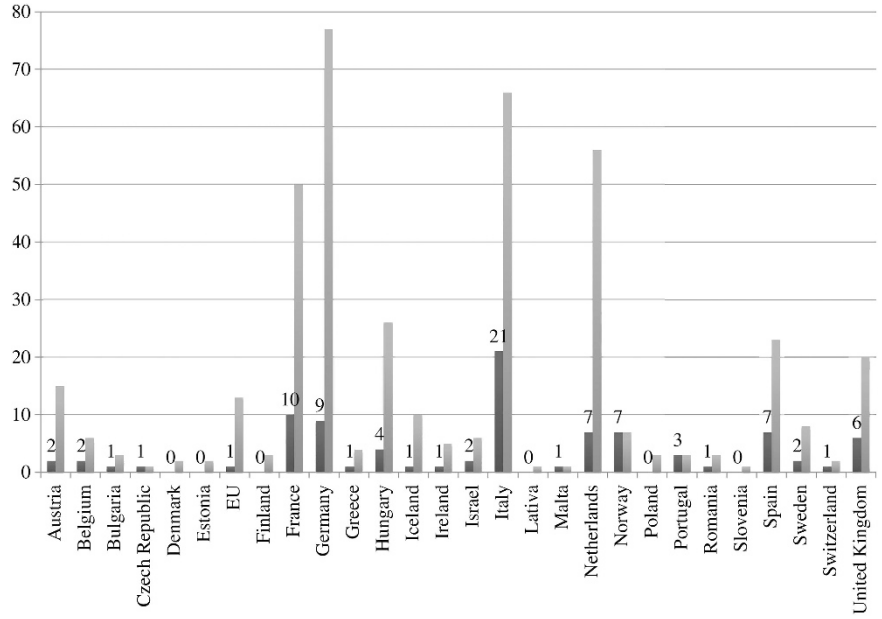

Figure 2 Country of biobank for European responders (dark grey bars with numerical labels) compared with the number of invited participants (light grey bars without labels).

Table 3 Results for [HUMAN] BIOBANK, SAMPLE and/or SPECIMEN, SAMPLE COLLECTION, STUDY and ALIQUOT

	European		*Swedish*		*All*	
Term and source	N	%	N	%	N	%
[HUMAN] BIOBANK						
BBMRI-ERIC	40	43	6	19	46	37
P^3G	33	36	16	52	49	40
MeSH	1	1	2	6	3	2
German Ethics Council	6	7	4	13	10	8
ISBER	8	9	3	10	11	9
Comment	4	4	0	0	4	3
SAMPLE and/or SPECIMEN						
OECD (2009)	25	27	7	23	32	26
P^3G	27	29	13	42	40	33
NCI	13	14	6	19	19	15
OECD (2006)	6	7	1	3	7	6
ISBER	21	23	3	10	24	20
Comment	0	0	1	3	1	1
SAMPLE COLLECTION						
Swedish Association of Local Authorities and Regions	43	47	14	45	57	46
Oxford English Dictionary	31	34	12	39	43	35
ISBER	17	18	5	16	22	18
Comment	1	1	0	0	1	1
STUDY						
MeSH	35	38	12	39	47	38
NCI	54	59	18	58	72	59
Comment	3	3	1	3	4	3
ALIQUOT						
P^3G	70	76	22	71	92	75
NCI	5	5	0	0	5	4
ISBER	17	18	8	26	25	20
Comment	0	0	1	3	1	1

Abbreviations: BBMRI-ERIC, Biobanking and Biomolecular Resources Research Infrastructure—European Research Infrastructure Consortium; MeSH, Medical Subject Headings; NCI, The National Cancer Institute Thesaurus.

Table 4 Results for CODED/CODING, IDENTIFYING INFORMATION/IDENTIFIABILITY, ANONYMISED/ANONYMISATION, PERSONAL DATA and INFORMED CONSENT

	European		*Swedish*		*All*	
Term and source	N	%	N	%	N	%
CODED/CODING						
OECD (2009)	47	51	13	42	60	49
P^3G	44	48	17	55	61	50
Comment	1	1	1	3	2	2
IDENTIFYING INFORMATION/IDENTIFIABILITY						
OECD (2009)	58	63	18	58	76	62
P^3G	28	30	11	35	39	32
ISBER	5	5	1	3	6	5
Comment	1	1	1	3	2	2
ANONYMISED/ANONYMISATION						
OECD (2009)	28	30	11	35	39	32
P^3G	62	67	19	61	81	66
Comment	2	2	1	3	3	2
PERSONAL DATA						
Directive 95/46/EC	38	41	21	68	59	48
P^3G	53	58	10	32	63	51
Comment	1	1	0	0	1	1
INFORMED CONSENT						
OECD (2009)	10	11	7	23	17	14
MeSH	11	12	2	6	13	11
P^3G	37	40	11	35	48	39
ISBER	34	37	10	32	44	36
Comment	0	0	1	3	1	1

Abbreviation: MeSH, Medical Subject Headings.

CODED/CODING

Of the two alternative definitions for CODED/CODING there was a tie between the definition from OECD[13] and P^3G[11] in the group comprising all respondents. One EU respondent commented that neither of the specified definitions were satisfactory, but also did not know of a better one. A Swedish respondent stated that the term CODED/CODING should be replaced with the term *pseudonymization*.

IDENTIFYING INFORMATION/IDENTIFIABILITY

For the term(s) IDENTIFYING INFORMATION/IDENTIFIABILITY, the definition by OECD[13] was the most popular among respondents. One European respondent stressed the difference between intentionally trying to identify a specific individual, and information linkage for a specific donor in order to create a valuable research asset but without any interest in revealing the identity of the donor. A Swedish respondent commented that the definition no. 3 (by ISBER[12]) corresponds to information that may directly or indirectly identify an individual, whereas definition no. 1 (by OECD[13]) is more related to a key that can be used to link the individual and data before and after pseudonymization. The same respondent also argued that the term IDENTIFIABILITY is something different than IDENTIFYING INFORMATION.

ANONYMISED/ANONYMISATION

Of the two given definitions for ANONYMISED/ANONYMISATION, the one given by P^3G[11] was favored by approximately two-thirds in both groups of respondents. Comments were provided by two European respondents, who stated that the definition by OECD[13] is the correct definition for ANONYMISED data, whereas the definition by P^3G is the correct definition for the process, and also that ANONYMISATION is the ability to identify the subject in terms of civil state from any type of measurements or combination of measurements has been lost. A Swedish respondent commented that in some Swedish basic legal documents the term is used for coded information.

PERSONAL DATA

The two given definitions for PERSONAL DATA were about equally favored, all respondents considered, with a small advantage for the definition given by P^3G.[11] There was, however, a considerable difference in the view of the definitions between the European group, who preferred the P^3G definition, whereas the Swedish group of respondents favored the definition given in the current European data protection directive.[10] One European respondent referred to earlier given comments and did not select a particular definition.

INFORMED CONSENT

For INFORMED CONSENT, all groups preferred the definition given by P^{3}G,[11] although the definition by ISBER[12] was almost as popular. One Swedish respondent pointed out that there might be a difference in the meaning of INFORMED CONSENT and *the decision of an* INFORMED CONSENT.

DISCUSSION

All in all, 123 persons participated in the survey. For European responders, the moderate response rate may be partially explained by 47 contacts who did not respond themselves but who had a responding co-contact with the same biobank affiliation. It is plausible that contacts connected to the same organization communicated and decided who should respond on behalf of their biobank, although the survey was indeed aimed to individuals rather than organizations. In addition, for European contacts, accounting not only for permanently undeliverable mails (that is, hard bounces), but also for so-called soft bounces ($N = 43$) caused by – for example, an overfull mail-box – will increase the European response rate to $\sim$25%.

The survey demonstrated variability in preference of definitions for most terms. In this section, we have analyzed this variability from different perspectives, guided by the quantitative result and comments from responders. We have aimed to compare the definitions by reasoning, while accounting for the outcome of the survey, and try to suggest how definitions may be improved.

Scope of definitions

At least four types of biobanks have previously been identified: (1) biobanks established as part of the health-care process; (2) biobanks established in the context of clinical trials; (3) biobanks comprising the samples collected in a specific research project and could be re-used for other research; and (4) population-based biobanks, which may have a more general research purpose.[4] Hence, it is desirable that a definition for the term [HUMAN] BIOBANK is general enough to contain all the four categories, in line with one comment that 'the clinical use' should be included in the definition. However, the most popular definition for [HUMAN] BIOBANK was the one given by P^{3}G, despite that the P^{3}G definition exclusively relates biobanks to population-based research. In a similar manner, the definition by ISBER for the term SAMPLE COLLECTION explicitly mentions research as a purpose. In contrast to SAMPLE COLLECTION, we argue that the term STUDY is firmly linked to a research question and may hence be thought of as a SAMPLE COLLECTION for which an ethical study permit exists.

With regards to the definition of what is intended as a 'study', the one from NCI was favored by respondents. If we can support this definition in the scope of clinical trials, where 'detailed examination' of a subject is the starting point to gather information and data, this definition is not really fitting with what is expected from biobanks in the sense that biobanks are mainly created as a resource aiming at contributing to various projects.[22] As a consequence, the scope and expected functions of informed consent could vary a lot from one design to another. If we can agree that informed consent is a process (see OECD definition) as it has to be continuous for the whole duration of the research program, it cannot be reduced to a simple procedure. That is why, the definition from P^{3}G, retained by respondents, is broader and is in accordance of what is expected from informed consent in the context of biobanks: expression of a will depending on the nature of the biobanks.

In the case of SAMPLE and/or SPECIMEN, the definition by P^{3}G may be challenged in popularity by the preference for OECD (2009) and ISBER combined. The latter two definitions differ only in three aspects for the SPECIMEN part: an addition of 'urine sample', and replacements of 'taken' with 'obtained', and 'subject or donor' with 'participant'. Hence, a combination of these definitions may be the preferable one.

Potential regional differences

The preferred definitions for the terms [HUMAN] BIOBANK, SAMPLE and/or SPECIMEN and PERSONAL DATA seem to differ to a larger degree between the European and Swedish groups than the rest of the terms. The P^{3}G definition for [HUMAN] BIOBANK was especially popular among Swedish respondents, which was also the reason that it scored highest among the total respondents. Contributing to the popularity of the definition by BBMRI-ERIC among European responders may be a higher awareness among European researchers about ongoing international infrastructure collaborations.

Differences in semantics

For the case PERSONAL DATA above, the two definitions are actually semantically different; the definition from the Directive 95/46/EC does not state that PERSONAL DATA *per se* lead to the identification of a natural person, only that it is the 'information relating to an identified or identifiable natural person'. In article 2.1.a of the Directive, personal is defined as 'any information relating to an identified or identifiable natural person; an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity'. Data that cannot be connected to an individual person therefore falls outside the scope of the Directive (Article 3 of the Directive). This current definition includes health data that are considered in addition as sensitive data.

The definition by P^{3}G, on the other hand, makes PERSONAL DATA synonymous with IDENTIFYING INFORMATION (see discussion above). We cannot be certain whether the responders noted this difference in semantics and reacted upon it or not. At least for European purposes we would at present, even if this might change, recommend to use the definition as stated in the Directive 95/46/EC, as to make biobank terminology consistent with legal terminology as far as possible. In the context of the revision of the Directive on Data protection (to be turned into an European regulation), the definition of Personal Data and sensitive Data (which will probably include Genetic Data) will be harmonized in all the European Union Members.[23] This will facilitate the common understanding of this terminology and will improve the communication between the research teams.

Three definitions were given for the terms IDENTIFYING INFORMATION/IDENTIFIABILITY, of which two, by OECD (2009) and ISBER, had been given in the context of the first term, whereas one, P^{3}G, was given in the context of the latter, see Table 2. The definitions were found to be semantically comparable, although use of the term IDENTIFIABILITY itself was questioned by one respondent. The OECD (2009)[16] definition was considered most popular among responders regardless of group. The definition by P^{3}G also brings up the concepts of CODING and pseudonymization. Although there is a relation between all these terms, it is possible that the inclusion makes the definition appear less straightforward in comparison. For the P^{3}G Lexicon, we propose that IDENTIFIABILITY is replaced with

IDENTIFYING INFORMATION and that the definition for PERSONAL DATA is used instead of the current one.

Definitions in the context of ontologies

The potential of an ontology for the biobank-administration domain has recently been described by Brochhausen *et al.*,[24] where the major benefit of an ontology in this context is presented as minimizing the effort of querying multiple databases for the same kind of samples of interest. The ontology, Ontologized Minimum Information About Biobank data Sharing (OMIABIS), uses the definition for [HUMAN] BIOBANK adapted from the German Ethics Council, see Table 1, which did only receive 8% of the total votes. This highlights an important difference between ontologies and terminologies: ontologies are designed to fulfill different requirements than terminologies. Therefore, they follow different design principles than terminologies.[25] Mainly, definitions in ontologies are written in a way that refers to the taxonomy underlying the ontology facilitating understanding by ontologists, and thus foster coordination of modular ontologies. Typically, definition should be authored following this patter: 'An *A* is a *B* with property *C*', where *A* are the entities defined, *B* is the immediate superclass and *C* is what makes the members of *A* different from all other members of *B*. This kind of definition is called Aristotelian definition.[26] Although Aristotelian definitions might not be intuitively descriptive for the domain experts, ontologies and the entities represented in them should be presented in a manner that is understandable to the aforementioned experts. To achieve that we suggest to, firstly, ensure coextensive reference for the favored definition with the definition provided by OMIABIS and secondly to add an annotation (an rfds:comment) containing the favored definition.

CONCLUSIONS

With the domestic and international proliferation of biobanks and their associated data, a common language for biobanks are essential. At present there is a considerable confusion in some of the terms used in the biobank community.

Indicative from the survey is the risk of focusing only on the research aspect of biobanking in definitions. By not also including the clinical area of application the likelihood of separated communities increases. Hence, it is the recommendation that important terms should be formulated in such a way that all areas of biobanking are covered, at least if the aim is to improve the bridges between research and clinical application. The generalizability of a term will of course depend on the scope of its definition. There is, however, nothing stopping us from using a hierarchical level to define different subclasses of the terms, and how they relate to different types of biobanks. Here, the semantic structure of an ontology will help.

In general, the outcome of this survey, which was mainly targeted at associated members of the European BBMRI, favors the glossary of the P^3G consortium whose definitions were voted most popular for seven of the eight terms where it was represented, all responders considered. The outcome of the survey should in the short run be accounted for by the related organizations whenever an update of their respective vocabularies is pending. With the risk of considering definitions out of their context, and only acknowledging the European perspective, the results could be used in the long run for the creation of a global biobank data dictionary, supporting information sharing about biological samples. In addition, the creation and maintenance of a machine-interpretable ontology representing the biobank domain would be beneficial.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We would like to thank all the responders who took their time to help us with this survey and the Swedish Research Council for granting the BBMRI.se project (grant agreement 829-2009-6285).

1 Shendure J, Ji H: Next-generation DNA sequencing. *Nat Biotechnol* 2008; **26**: 1135–1145.
2 Martin JA, Wang Z: Next-generation transcriptome assembly. *Nat Rev Genet* 2011; **12**: 671–682.
3 Yuille M, van Ommen GJ, Brechot C *et al*: Biobanking for Europe. *Brief Bioinform* 2008; **9**: 14–24.
4 Harris JR, Burton P, Knoppers BM *et al*: Toward a roadmap in global biobanking for health. *Eur J Hum Genet* 2012; **20**: 1105–1111.
5 Wichmann HE, Kuhn KA, Waldenberger M *et al*: Comprehensive catalog of European biobanks. *Nat Biotechnol* 2011; **29**: 795–797.
6 Norlin L, Fransson MN, Eriksson M *et al*: A minimum data set for sharing biobank samples, information, and data: MIABIS. *Biopreserv Biobank* 2012; **10**: 343–348.
7 Knoppers BM, Saginur M: The Babel of genetic data terminology. *Nat Biotechnol* 2005; **23**: 925–927.
8 Rogers FB: Medical subject headings. *Bull Med Libr Assoc* 1963; **51**: 114–116.
9 Sioutos N, Sd Coronado, Haber MW *et al*: A semantic model integrating cancer-related clinical and molecular information. *J Biomed Inform* 2007; **40**: 30–43.
10 The European Parliament and the Council of the European Union (1995). Directive 95/46/EC of the European Parliament and of the Council. Available. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31995L0046:EN:NOTcurrently under revision, to be turned into a Regulation. http://ec.europa.eu/justice/data-protection/. Accessed 10 December 2013.
11 Public Population Project in Genomics and Society. Lexicon. Available at. http://www.p3gobservatory.org/lexicon/list.htm. Accessed on 20 February 2013
12 International Society for Biological and Environmental Repositories (ISBER). Collection, storage, retrieval and distribution of biological materials for research. *Cell Preserv Technol* 2008; **6**: 3–58.
13 Organisation for Economic Co-operation and Development (OECD) (2009) OECD Guidelines for Human Biobanks and Genetic Research Databases. Paris. Available at. http://www.oecd.org/science/biotech/44054609.pdf. Accessed on 20 February 2013
14 Organisation for Economic Co-operation and Development (OECD) (2006) Creation and Governance of Human Genetic Research Databases. Paris.
15 Biobanking and Biomolecular Resources Research Infrastructure (BBMRI) (2012) Statutes for the Biobanking and Biomolecular Resources Research Infrastructure - European Research Infrastructure Consortium (BBMRI-ERIC). Available at. http://www.meduni07.edis.at/files/draft_bbmri_statutes_clean.pdf. Accessed on 20 February 2013.
16 Deutscher Ethikrat (2010) Human biobanks for research. Berlin. Available at. http://www.ethikrat.org/files/der_opinion_human-biobanks.pdf. Accessed on 2 April 2013.
17 Swedish Association of Local Authorities and Regions (2011) Glossary (In Swedish). Available at. http://www.biobanksverige.se/getDocument.aspx?id=50. Accessed on 20 February 2013.
18 Oxford English Dictionary (2013) Available at. http://www.oed.com/viewdictionaryentry/Entry/36275. Accessed on 20 February 2013.
19 EuroBioBank. Available at. http://www.eurobiobank.org/en/information/glossary.htm. Accessed on 15 January 2013
20 Riegman PHJ, Morente MM, Betsou F, de Blasio P, Geary P: Biobanking for better healthcare. *Mol Oncol* 2008; **2**: 213–222.
21 Norweigan Ministry of Health and Care Services (2008) Lov om medisinsk og helsefaglig forskning. Available at. http://www.lovdata.no/all/tl-20080620-044-001.html#4. Accessed on 15 January 2013.
22 Expert group of the European Commission (2012) Biobanks for Europe, a challenge for governance. Available at. http://ec.europa.eu/research/science-society/document_library/pdf_06/biobanks-for-europe_en.pdf. Accessed on 25 June 2013.
23 The European Commission (2012) Proposal for a new regulation COM(2012) 11 final, 2012/0011 (COD). Accessed on 25 June 2013. To follow up on the revision process see. http://ec.europa.eu/justice/data-protection/
24 Brochhausen M, Fransson MN, Kanaskar NV *et al*: Developing a semantically rich ontology for the biobank-administration domain. *J Biomed Semantics* 2013; **4**: 23.
25 Smith B: From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform* 2006; **39**: 288–298.
26 Smith B, Ceusters W: Ontological realism: a methodology for coordinated evolution of scientific ontologies. *Appl Ontol* 2010; **5**: 139–188.